DOCUMENT RESUME

ED 075 254	24	SE 016 083

AUTHOR	Branson, Robert K.; And Others
TITLE	Self-Paced Physics, Documentation Report, Final Report 5.0.
INSTITUTION	Naval Academy, Annapolis, Md.; New York Inst. of Tech., Old Westbury.
SPONS AGENCY	Office of Education (DHEW), Washington, D.C. Bureau of Research.
BUREAU NO	BR-8-0446
PUB DATE	71
CONTRACT	N00600-68-C-0749
NOTE	108p.
EDRS PRICE	MF-$0.65 HC-$6.58
DESCRIPTORS	*Academic Performance; College Science; *Curriculum Development; *Physics; Program Descriptions; *Program Evaluation; Science Education; Self Help Programs; *Statistical Analysis; Study Guides
IDENTIFIERS	Self Paced Instruction

ABSTRACT

As a supplement to the principal reports, descriptions are given in this report for the development, validation, and installation of the Self-Paced Physics Course at the U. S. Naval Academy. Following an executive summary, an introduction to course characteristics, and an overview of the project, statistical tests are discussed in connection with discriminant analysis, one-way analysis of variance, and step-wise regression. Sample data from one experimental group and two control groups, collected during the Fall 1969 tryout from weekly posttests, final examinations, and reported proctor time, are used together with background scores in statistical calculations. Relationships are determined between audiovisual and non-audiovisual groups and among the variables, and student performance including differences resulting from individual test items is studied in relation to study guides. Learning category, confidence, and difficulty ratings are also analyzed. As a result claimed in Technical Report 5.6, this multi-media course is at least as effective as traditional courses. Besides 53 tables, the formula used in discriminant analysis and the learning category taxonomy in a problem form are given in the appendices. (Related documents are SE 016 065 - SE 01 088 and ED 062 123 - ED 062 125.) (CC)

ERIC

Set #3

ED 075254

U S DEPARTMENT OF HEALTH
EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY

DOCUMENTATION REPORT

UNITED STATES NAVAL ACADEMY

Self-Paced

PHYSICS

This document is a supplement to the principal reports 5.10, 5.9, and 5.8, developed and produced under the U. S. Office of Education, Bureau of Research Project #8-0446, for the U. S. Naval Academy at Annapolis, Maryland. Contract #N00600-68C-0749.

5.0 FINAL REPORT

016 083

FILMED FROM BEST AVAILABLE COPY

ED 075254

SELF-PACED PHYSICS

FINAL REPORT

Submitted by New York Institute of Technology

Old Westbury, New York

Prepared by

Robert K. Branson
James K. Brewer
William A. Deterline

Developed and produced under the U.S. Office of Education, Bureau of Research Project #8-0446, for the U.S. Naval Academy at Annapolis Contract #N00600-68C-0749

Technical Report
5.0

ERIC

TABLE OF CONTENTS

I. EXECUTIVE SUMMARY

This Final Report, and the accompanying documents listed on page 8 , describe the physics course delivered under contract Number N00600-68C-0749, its development, validation, and installation at the U. S. Naval Academy. The materials to be used in the course are separately packaged and are described in the letter of transmittal.

The course, as delivered, is self-paced, independent study, multimedia, computer or manually managed, classical introductory physics. It is completely packaged and can be used at the U. S. Naval Academy with any number of midshipmen, in the fleet, or, at any other place having a need for the content contained in the objectives as listed in Technical Reports 5.1, 5.2, and 5.3, provided that adequate faculty support is available at student request.

A second major component of the delivery is the Empirical Course Development Model which sets forth the procedures for developing new courses or adding content to existing courses. Capable professionals, through the use of the model, can design and develop self-optimizing courses or segments of courses.

A third major component of the delivery is the report on the research, evaluation, and validation procedures followed in bringing the course to completion. This information is contained principally in Technical Report 5.0: Final Report.

The course is fully operational at the present time. the Physics Faculty can decide, at any time, to offer the

independent study course, or to continue the present procedure of offering the students a choice between the new course and the traditional course. It is the contractor's recommendation that this decision be reached by the Naval Academy, taking all relevant factors of current operations into consideration.

The benefits to be gained by offering the independent study version should be in the ability to schedule the course on a flexible basis at all hours during the day thus avoiding conflicts with other courses. Further, through time, a gradual reduction in the number of people required to teach the course should be realized, if this reduction should be desirable to the Academy.

Because of the background and capabilities of the midshipmen, we have concentrated on print media and de-emphasized the more complex and expensive audiovisual approach. The Empirical Course Development Model does, however, outline a procedure for developing, using, and evaluating available audiovisual media. The basis for the decision is presented in this report, particularly in the results and discussion, and in the Revision Process Documentation.

The fact that the systems approach to course development has been effectively employed in this course should be carefully considered in analyzing future needs. The contractor believes that an extension of the systems approach to other courses and programs at the Academy would have benefits well above those possible to achieve in a physics course alone. Particularly, in the Effectiveness Report, the point has been

made that an absolute standard does not exist for the amount of physics that should be learned in this course, and there is no specification of the level of mastery expected. Further study of this question should yield a definition of physics knowledge requirements for students with any Academy major.

Finally, the installation and operation of the course at the Academy during 1970-71 should represent only the beginning of the optimizing process. The course is designed to iterate and to be systematically revised. In this feature lies the power of the development model.

The course is completed. It is quite effective. It is installed and operating at the Academy. Additions are being made to it according to the procedures recommended, and it should be revised and improved following each iteration. Management attention should be directed toward seeing that the continuous improvements are made, that course operations are adequately supported with staff or computer assistance, and that staff levels for course operations are consistent with overall needs of the Physic Department and the Naval Academy.

II. INTRODUCTION

The Purpose of the Contract

Exerpting from the original Request for Proposal (RFP), the objectives of the contract were, "...to develop, test, and evaluate the best possible instructional media, materials, and strategies, utilizing all available techniques in the current state of the art." Within the context of this objective and listed later in the RFP are requirements for tryouts, revisions, and evaluations of the various media and materials with the intent of including those materials which tend to optimize the course results and eliminating those materials which do not appear to contribute to student performance.

In any research and developmental effort of the magnitude of this contract requiring state-of-the-art technology there are generally three sources of information which emerge as relevant, and from which one can infer the ultimate intent of the contract. The first source includes the RFP and initial contract documents. The second source is the modified contract and official agreements that have been worked out between the contractor and the customer. The third represents the formal and tacit agreements reached between contractor and customer staffs as they progressed toward the delivery of the final system. These sources of information will be discussed in order and hopefully will provide the basis for establishing clearly the intent of the contract as interpreted both by the contractor and the customer.

Throughout the RFP, the requirement for design, development, tryout, and revision of instructional materials is reiterated. In addition to the instructional materials, various approaches to the scheduling and operation of the course had to be worked out so that the finally delivered package included not only a set of effective[1] materials but also a set of procedures which will allow the Academy, in future iterations of the course, by evaluating and revising, ultimately to achieve an optimum course based on the original version and techniques applied by the contractor. The Empirical Course Development Model (TR 5.7) sets forth the recommended procedures for future revisions.

Thus, two important considerations are established, (1) that the course must be developed according to an empirical methodology, and, (2) that the final implementation package should include techniques for reiteration of the course with successive improvements to be made by the faculty in the future.

Since the term, "best possible" is not officially defined in the contract or the RFP it was assumed that the meaning of this concept should be derived from the contractor's analysis and approach to the problem so that the best possible finally delivered course would include not only the instructional media, materials, and strategies, but would also take into consideration the realistic constraints of the Naval Academy.

---

[1]See Technical Report 5.6: Effectiveness Report for a thorough definition and discussion of "effectiveness."

Such constraints, likely to be found in any university, include the judgment of faculty members based on their knowledge of content of other courses, the extra-curricular activities at the Academy, and the long range goals and objectives of the Academy.

During the materials development effort for the initial tryout, a number of discoveries was made by the contractor which resulted in design changes and allocations of time and resources for the computer. While it was initially considered to make the computer the central feature of the entire course, the results of outside investigations and recent experience at the Academy suggested that a redefinition of the computer's role was imperative if the course were going to be used successfully. It was agreed that the computer was best used as a management tool (Computer Managed Instruction) rather than an instructional device, except for a specific application of computers in solving physics problems with many variables.

The third source of specifications and agreements on the final package emerged as contractor and Academy staffs cooperated to implement the interim versions. It was during these tryouts that the various possible roles of instructors, amount of dependence on the computer, the function of the laboratory, and the overall operational methodology of the course were worked out to fit Academy requirements.

Several additional reports have been prepared to be submitted in conjunction with and as a part of this Final Report which speak to various issues and aspects of the course as developed and finally submitted.

Documentation Reports

1. FINAL REPORT (TECHNICAL REPORT 5.0)

A. Description of the methods, activities, materials and reports developed and produced by the contractor to satisfy the requirements of contract N00600-68C-0749, and of the Request for Proposal as a part thereof.

B. Statistical data and statistical tests where appropriate, and, the conclusions drawn and recommendations made by NYIT on the basis of the analysis of these data and of the experience gained during the project.

C. Specific reference to the Evaluation and Validation Design (Technical Report 4.7) and an interpretation of the data collected as specified by the Evaluation and Validation Design.

D. Based on the foregoing empirical data and experience, revisions to the Design for the Selection of Strategies and Media (Technical Report 4.9). These recommendations and revisions are contained in Technical Reports 5.1, 5.2.1, 5.2.2, 5.3, 5.4, and 5.5.

2. COURSE DESCRIPTION (TECHNICAL REPORT 5.1)

The course as delivered can be used effectively as an independent study, self-paced computer managed introductory physics course. The course materials can be used as supplements to traditional instructional techniques, or through appropriate management and staff assignment procedures, the course can be used to increase the number of students taught by each qualified instructor. Particular configuration used

should depend upon the needs of the Naval Academy at any given point in time.

3. COURSE OBJECTIVES (TECHNICAL REPORT 5.2.1)

Each of the performance objectives is represented by a problem so that the level, scope, and assessment measures are described in unambiguous form. The principal requirement for student success is that he be able to work the problems appropriately in the time allowed.

4. COURSE STRUCTURE AND SEQUENCE (TECHNICAL REPORT 5.5.5)

The topical sequence of objectives including the decision processes which led to this sequence.

5. TEST ITEM BANK (TECHNICAL REPORT 5.3)

A compilation of criterion check items and diagnostic test items identified by terminal objectives. The item bank includes multiple questions for each terminal objective and item statistics collected during the tryout conducted in the Fall of 1969.

6. MANAGEMENT SYSTEM REPORT (TECHNICAL REPORT 5.4)

A description of course implementation procedures recommended by the contractor, the nature and form of the test, the method of scoring and recording scores, the kinds of feedback provided the students, and the method of presenting the feedback. A description of all record-keeping procedures and the forms on which records can be kept. The report details the use of the computer managed instruction system where applicable.

7. REVISION PROCESS DOCUMENTATION (TECHNICAL REPORT 5.5)

A description of the specific empirical revision activities,

rationale for these activities, and a compilation of the data upon which revision decision were made.

8. EFFECTIVENESS REPORT (TECHNICAL REPORT 5.6)

The effectiveness report discusses effectiveness first and primarily in terms of "mission" effectiveness, by explaining the procedure used to derive the mission for the course, by elaborating a definition of effectiveness in terms of this derived mission, and an explanation of how the learning system developed in this contract achieves the derived mission.

9. EMPIRICAL COURSE DEVELOPMENT MODEL (TECHNICAL REPORT 5.7)

This report presents the final version of the Empirical Course Development Model, expanded to include additional techniques so that the model will be applicable to a broad range of courses in instructional settings.

The Empirical Course Development Model is presented as a complete set of procedural guides, with supplementary devisions and explanations. The model was designed, first, to provide the Naval Academy Physics Department with the tools and procedures necessary for continued course optimization, and, second, to furnish these same tools to the Academy and other schools and colleges for empirical course development in any subject matter area.

The Empirical Course Development Model is intended to be used in conjunction with the Management System Report, Technical Report 5.4, and the Effectiveness Report, Technical

Report 5.6

These repo ively, are submitted with the intention of an least these funamental questions:

1. How did the contractor fulfill the terms and conditions of the contract as amended?

2. What materials and level of detail are necessary in order for the Physics Department to understand the development process as it was practiced by the contractor, and, what specifications procedures and plans are necessary to allow the Physics Department to continue course development as recommended by the contractor?

We believe that the reports as submitted are sufficient to achieve both purposes, without containing unnecessary detail and recommendations beyond the scope or training of the Physics Department.

ERIC

III. PROJECT DESCRIPTION

Overview

There were three imp tant milestones in the development of this project. The first was the development of the content and objectives of the entire course, the second was the major tryout in the Fall of 1969 of all of the learning materials at the Academy, and the third was the implementation of the course as it was intended to operate in the Spring and Fall semesters of 1970. The initial tryouts of the materials at the Academy occurred during the last part of the Fall of 1968 and during the Spring of 1969. The Fall 1968 materials were the first rough draft version and were used with only a few students in order to determine level of expectation, quality of materials and time requirements. The second tryout involved a considerably greater amount of material and lasted for the entire Spring 1969 semester with approximately one hundred students. On the basis of these two early iterations of the materials the configuration of the course in the Fall of 1969 was designed.

The purpose of the Fall 1969 tryput was to compare students preference for, and, the performance of the instructional materials selected and packaged for the course. On the basis of performance and student preference data, combined with judgments about costs and general effectiveness, the combination package of the final version of the course was to be derived. After the Fall 1969 tryouts, changes took place in the method of course operation, and the course content.

Further, Academy faculty had the first opportunity to develop materials according to the Empirical Development Model. The faculty materials development effort was concerned principally with adding content to the original course in order to meet newly prescribed Academy curriculum requirements.

Extensive data was collected during each iteration of the course for the purposes of making judgments about materials adequacy and for providing information upon which revision decisions could be based. Several large boxes of computer printouts, including individual student response records, synopses of group response records, performance data, rating data, and time data have been collected and are available for detailed review. It should be emphasized that these materials are not useful for summative evaluation. The iterations during the Fall of 1969 and the Spring of 1970 provide data useful to ascertain total value of the course. Specific conclusions and recommendations about the course and its operations are made in Empirical Course Development Model (TR 5.7) and in the Effectiveness Report (TR 5.6).

The data presented in this report were collected during the Fall 1969 tryout.

The Fall, 1969 Tryout

The project design led up to the Fall 1969 tryout as the major data source for materials revision purposes and to the Spring 1970 tryout for the final data source for management decisions. In the Fall of 1969, midshipmen were required to complete the entire sequence of instruction, regardless of

whether they were capable of completing the course early. This procedure was followed in order to retain the complete Academy range of talent in the samples. During the Spring of 1970, midshipmen were allowed to complete the course early, provided that they passed stringent criteria, in which case they were exempted from the final examination.

The major Technical Reports, submitted in 1969, describe in detail the objective considerations and rationale for the Fall 1969 tryout. These reports are:

1. Technical Report 4.7: Rationale for Sequencing Objectives. Finkel, 1969.

2. Technical Report 4.7.1: Evaluation and Validation Design. Deterline and Branson, 1969.

3. Technical Report 4.7.2 The Validation Process. Deterline and Branson, 1969.

4. Technical Report 4.9: Design for Selection of Strategies and Media. Deterline and Branson, 1969.

5. Technical Report 4.3: Course Revision and Restructure. Vierling, 1969.

6. Technical Report 4.12: Weekly Course Segment Documentation, Weeks A through O.

Description of the Course

The version of the physics course delivered to the Academy includes instructional materials, a management system, and, recommended evaluation procedures. The instructional materials were packaged to be used within the recommended management system according to the procedures described in Technical Reports 5.1, 5.4., and 5.7. The configuration of the course is independent study with faculty management and support for personal consultation upon request by the student. The instructional materials consist of the basic textbooks, the problems and solution books, the criterion test items, and the selected audiovisual materials.

The recommended configuration of the course and the basis for it, are completely responsive to the RFP, particularly in terms of the explicit objectives set forth on page one: "The objectives of the program are to develop test and evaluate the best possible instructional media, materials, and strategies, utilizing all available techniques in the current state-of-the-art." Elsewhere in the RFP specific reference is made to Computer Assisted Instruction (CAI) and Computer Managed Instruction (CMI) and the utilization of appropriate instructional media, particularly audiovisual materials.

On the basis of the data available when the RFP was issued, there were excellent reasons to believe that specific audiovisual materials, Computer Assisted Instruction, and Computer Managed Instruction would all contribute uniquely to the performance of midshipmen in the physics course. Our general results, as

well as results reported by other investigators, seem to indicate that the differential effects of media in experiments on highly verbal students are rarely convincingly demonstrated. The contractor, therefore, has taken the view that nothing should be included in the course which added either to student time or to course expense unless a clear demonstration of its utility could be made. The statistical analyses, results, and discussion sections present the data and conclusions reached by the contractor after careful evaluation of the media and course structure.

The report of an exhaustive search of the literature by Dubin and Taveggia (1968) confirm the difficulty of demonstrating statistically significant differences in teaching methods at the college level. Indeed, if it were not for the finding that time spent in study did contribute to improved grades for college students, their analysis would have yielded no positive results at all. Their findings do not mean that students do not learn, but, simply that they learn about as well under virtually any teaching method.

IV. STATISTICAL ANALYSES

Introduction

In this section analyses are described testing population hypotheses using samples of scores from weekly posttests (both total correct and log confidence[1] scores), the final examination and reported proctor time. It is assumed throughout, with the usual misgivings, that the underlying assumptions are valid for conducting the statistical tests and procedures. In those procedures where the calculation of the power of the test is appropriate the formula

$$N = \left[(z_\alpha + z_\beta)\sigma/d \right]^2$$

has been used to calculate the power.

Power here, as in all statistical discussions, refers to the probability of rejecting the null hypothesis when in fact it is false. It is desirable that power be as close to 1 as possible and this generally requires that the sample size in the analysis be quite large. Since the samples used were large the obtained power value were above .99 for $d = .5\sigma$ Cohen (1969) suggests that a power .80 is sufficient for most behavioral studies. The symbol "d" is effect size and reflects how much difference one is willing to tolerate before declaring a significant difference exists. If one desires to detect very small differences then d is small and large samples are generally necessary. The reverse is true for detecting large differences, i.e., d large. Throughout this

---

[1] The log confidence scoring procedure is detailed in Technical Report 4.7: Evaluation and Validation Design, page 28ff.

paper d was set at $.5\sigma$. Cohen (1969) calls this a "medium" level. The symbols z_α and z_β denote the standard normal deviate values (two-tailed) for α and β values respectively. Unless otherwise noted, * and ** will respectively denote significance at the .05 and .01 levels with α set at .01 or smaller.

It cannot be over emphasized that the reader must reach his own conclusions as a result of the statistical tests in light of the stringent assumptions of the various tests. These basic assumptions will be given prior to each distinct set of analyses.

Discriminant Analyses

In order to utilize information on individual differences in predicting the potential behaviors of a student, the method of classification by discriminant analysis was selected. A discriminant analysis compares an individual profile with that of a group and calculates the probability of membership in that group. The details of and tables for calculation of the probabilities of group membership for several variables are given in Appendix A. All calculations were made using the BMD05m Computer program.

The grouping variables here are primarily of two types: performance and efficiency, with high, medium, and low levels defined for each. The variables which comprise the profile for each individual are five Strong scores (ACH, M-F, OCL, SIN and SPL), one Naval officer score (denoted NAV), SAT (V), SAT (M), High School rank (converted), Whole Man Score,

Quality Point Rating (QPR) for the third semester and the Physics Validation Score. These twelve variables and a subset of them were used as descriptors of individuals to provide a relatively wide range of background information for classification purposes. The reduced set of variables used were the result of a step-wise regression analysis on final examination scores with SAT (M), QPR, and Physics Validation composing the set.

To make the technique more applicable, probability functions are presented for each group so that an instructor or counselor can obtain an individual's background scores, substitute performance on final examination or medium level of time spent in a a particular week for that individual. (See Appendix A for a detailed explanation.) The technique: (1) assures the user that the errors of misclassifying each individual have been minimized; (2) provides a check on the accuracy of the classification when used with similar individuals; and, (3) gives a test of significance between groups over all variables.

The variables used were assumed to be distributed as multivariate normal within each group and were such that the covariance matrix was the same for all groups. The validity of the following analyses rests on these assumptions.

The norm referenced nature of most of the background scores gives some assurance of distribution normality (at least symmetry) but little can be said about equality of the covariances. What the user of these techniques has to do is remind himself that, "given the assumptions are valid then

this is how we may proceed." How "good" it is has to be determined empirically.

For a technical description and discussion of the technique of discriminant analyses see the references on Anderson (1958) and Cooley and Lohnes (1962).

Final Examination

Three types of students took the final examination.

1. Experimental: those who were exposed to the experimental treatments and media and took weekly pre and posttests.

2. Control I: those who took only the pre and posttests each week.

3. Control II: those who took neither pre nor posttests nor were exposed to any experimental condition or media.

Discriminant analyses were conducted for each of the types (1 and 3) above, the latter being denoted the Control Group.

The raw final score for each student was used as the performance measure and the three levels of performance were:

a) High: those scores which were greater than or equal to $\bar{x}$ + .5S. Under a normal distribution this includes approximately the upper 31% of the scores.

b) Medium: those scores which were absolutely between $\bar{x}$ + .5S and $\bar{x}$ - .5S. Under a normal distribution this includes approximately 38% of the scores.

c) Low: those scores which were less than or equal to $\bar{x}$ - .5S. Under a normal distribution this includes approximately the lower 31% of the scores.

Weekly Posttest Performance

For each of the weeks I through O (last 7 weeks)

discriminant functions were developed using the average log confidence score per student per week as the performance measure with the levels being defined as the final examination.

Time in Media (Efficiency)

Using the proctor's report of time (in minutes) spent in each student in each experimental condition as an efficiency measure, High Medium, and Low time groups were defined as above. Discriminant functions were developed which classified each student as being high, medium, or low in time given an experimental condition and background variables.

The following tables are presented for those discriminant analyses for which the group equality test was significant for $\alpha = .05$ or .01.

Tables 1, 3, 5, 7, 9, and 11 identify the performance grouping variable, profile variables, sample sizes and means for each profile variable by group for the previously described analyses.

Tables 2, 4, 6, 8, 10, and 12 display the classification check matrices giving the number of correct and incorrect classifications made by the discriminant functions for the samples used.

For example, in Table 2, the discriminant analysis classification procedure used correctly predicted 44 of 58 low performing control group students on the final exam given only their profile scores, 27 of 54 medium level students and 37 of 54 high level students. Chi-square analyses could be conducted on each of Tables 2, 4, 6, 8, 10, and 12, but this adds

very little if anything to the usefulness of the procedure. It takes no statistical test to conclude that the classification procedure is better than guessing.

Table 13 gives the generalized Mahalanobis D^2 statistic values to test the hypothesis that the mean values are the same for all groups for the profile variables (See Cooley and Lohnes, 1962). Subject identification, grouping variable, number of profile variables and degrees of freedom are also given in Table 13.

The classification check matrices demonstrate that the proportion of correct classification was generally above .50 and considerably higher than this for high and low performance groups.

The implication here is that the procedure is considerably better than guessing and would be of some value in obtaining an estimate of level of success on the final examination, given background scores. This procedure should assist in planning remedials, expected tutorial time and counseling. The object of the instruction would be to "beat the classification" and bring every student to criterion performance.

A comparison of actual outcome with predicted classification for each student would be a good indicator of attainment of this goal of instruction. For example, if a student has a high probability of being in some low performance groups on final examination then proper steps could be taken to overcome his deficiencies and bring him up to criteria on the final. Those students, however, who have a high

probability of being high performers could be routed through supplemental materials on topics which are easy for them or be used in a tutorial fashion with other, weaker students. Of necessity, the final examination would require procedural change if this approach were taken. Absolute criteria of performance would replace the relative standing criterion currently used. See Technical Report 4.7 page 53ff for a more complete discussion of this problem.

The result, by use of step-wise regression analysis, that SAT (M), QPR and Physics Validation were the best predictor of final examination performance is no surprise since these same type variables are usually best predictors of performance, especially QPR.

It is apparent that these discriminant analyses do not identify the best predictors but only indicate what is most probable, given background variable of several types. Of the 18 discriminant analyses performed only seven had D^2's sufficiently large to conclude that the groups (high, medium, low) had significantly different population means on the background variables. None of the proctor time (efficiency) analyses was significant, but this is conceivable since one would not expect drastically different background characterization for high, medium, or low reported time spent.

The reader will note that the use of twelve background variables for classification of experimental subjects on final examination performance (Table 4) was only slightly more accurate than the classifications using the three best

predictor variables (Table 8). For relatively rough screening and predictions on final examination performance it would appear that this reduced set of variables, representing considerable savings in testing times, and calculations would be quite satisfactory; we find the same variables occuring as best predictors in other weekly posttest analyses which will be discussed in detail later.

Weekly Posttest Comparisons

The samples of scores for the last seven weeks come from the experimental subjects only and the variables are total correct and log confidence averages on both media relevant and media non-relevant items. The analyses consist of step-wise regression each week, one-way analyses of variance with weeks as the treatment, randomized block designs (with weeks as blocks and media as treatments) and one-way analyses of variance with media as the treatment. Each will be discussed separately even though some analyses are closely related in rationale and results.

Regression Analyses

For each week and for each of the two variables, total correct and average log confidence score, a linear regression analysis was conducted to find the best set of predictors (among the 12 background variables) of performance for each week. The mathematical model used is the one proposed by Draper and Smith (1966) and has assumptions of homogeneity of variance, linearity, and normality of the error distributions. The analyses were conducted using the BMD02R computer

program.

A summary of the results of these analyses along with the multiple R values for each is given in Table 14. The analyses were conducted on the set of 77 experimental subjects who had complete data on both variables each week. The value F = 3.0 was used for both inclusion and exclusion concerning the 12 variables. (See Draper and Smith (1966), for explanations and discussions of the multiple regression technique).

Multiple R is used as an indicator of how good the predictions of performance would be if one used the variables resulting from the analysis. A value close to 1 is most desirable.

From Table 14 it is apparent that some of the same variables occur again and again as best predictors of weekly performance and include the same best predictors in general as were found for predicting final exam performance in the Discriminant Analysis section. The most common is QPR which occurs in 10 out of 14 analyses as a best predictor. It occurs five times as the best single predictor. It is worthy of note that SAT (V) and M-F occur several times but did not occur as predictors of final exam performance and that High School Rank did not occur for any week and SAT (M) occurred in only one week and [illegible] variable only.

If one were pressed to pick a best single predictor of performance on weekly posttests, he could probably do no better than to pick QPR and this would agree with the general findings of other researchers using similar performance criteria. This should be cautiously done, however, since

Table 14 shows that the percentage of total variance accounted for using all the best predictors range from 9.6% ($.31^2$) to 33.6% ($.58^2$) with the best variance proportion for QPR alone being 21.2% ($.46^2$).

These low percentage values indicate that using QPR alone will not yield very accurate predictions of performance even though under the circumstances it is the best from an empirical and theoretical standpoint. In other words, even the best predictor of performance is a poor one.

One-Way ANOVAS (Weeks as Treatments)

The basic assumptions for this set of analyses is that the samples are indepently drawn from treatment population which are normally distributed with equal variances. Since each week consisted generally of the same group of subjects, there is a serious question of sample independence. However, if one is willing to assume that the weekly samples, representing, as they do, different subject matter, are, for all practical purposes independent, then, this relatively important assumtion can at least be listed as "questionable". (See Hicks (1964) for details of assumptions and procedure). Alternative procedures which do offset some of these assumptions yielded the same general findings.

The items on which the scores are given consist of two types

1. Media Relevant items which are specifically designed to measure achievement of the Terminal Objectives in the

co[illegible] or which parallel media were designed. Those items ar[illegible]ented in Technical Report 5.3: Test Item Bank, along with [illegible]atistical description of them. They are related to the [illegible]al Objectives in Technical Report 5.5: Revision Process [illegible]ocumentation. The rationale for selection of the spec[illegible]TO's is presented in Technical Report 4.9: Design for [illegible]election of Strategies and Media. There was a total of [illegible] media relevant items spread across the fourteen weeks [illegible] the semester.

2. M[illegible] non-Relevant items which are specifically designed to me[illegible]e achievement of TO's on which no parallel media were devel[illegible]. The instruction sources for these items were the textbo[illegible], Study Guides, lectures, and laboratories.

B[illegible] sets of items represent subject matter which changes on a [illegible]ly basis. Tables 15 through 20, as explained later, sh[illegible] the weeks are different, probably because some of the [illegible]ics are more difficult.

The one-way analyses were conducted on these item types separately as well as pooled and BMD01V was used to conduct these analyses. Sample sizes in these and subsequent analyses vary due to class attendance differences.

Tables 15 and 16 show the results of the one-way analyses with weeks as treatments using scores only on the media relevant items.

T[illegible]es 17 and 18 show the results of the one-way analyses using scores only on the media nonrelevant items while tables 19 and 20 show the results for the pooled items scores.

The results shown in Tables 15 through 20 imply that weeks are different in mean score when measured on total correct and log confidence average using media relevant or nonrelevant items as well as pooled items.

Randomized Block Design Analysis

If one can assume that weeks can be treated as internally homogeneous blocks and media can be defined as treatments, then the scores within each media group can be defined as observations of a randomized block design. The purpose here is to test for equality of the media means with the restriction on randomization being the weeks. The basic idea is to offset the effect, in comparing media, that weeks are different. The BMD05V program was used for this analysis to handle the unequal sample sizes in the cells. (Hicks, (1964), gives a detailed discussion of this statistical technique).

Table 21 shows the results of this analysis for total correct over all items (both media relevant and media non-relevant). Table 22 shows the results for log confidence average for the same items. The assumptions are basically the same as for the one-way ANOVAS in the preceding section.

Of note is the significance between media means when weeks are defined as a restriction on randomization. The implication is that the media are different when compared within a week, knowing that weeks are different.

One-Way Analyses (Media as Treatments)

These analyses and their assumptions are basically

ERIC

identical to those above with weeks as treatments with the exception that the media now are being defined as treatments. The independence assumption is slightly easier to make in these analyses since each media group is composed of different subjects.

Tables 23 and 24 show the results for media relevant items, Tables 25 and 26 for media nonrelevant and Tables 27 and 28 for pooled items over all weeks.

Analyses identical to those in Tables 27 and 28 were computed within each of the weeks I through 0; Tables 29 through 42 show the results of these analyses for both variables on pooled items.

It is worthy of note that only for week L does the media show up as significant whereas when weeks is taken as a restriction on randomization as in the randomized block design, media shows up as significant.

Final Examination

The final examination in the course consisted of a 60-item test containing four 15-item subtests. The subtests were composed of constructed response items selected by two groups of instructors, and, multiple choice items selected by two groups of instructors.

There were three basic questions to be tested on the final examination:

1. Will experimental group mean performance exceed control group mean performance on the total test?

2. Will either of these two groups do better on multiple choice items than on constructed response items?

3. Will the variance of the experimental group be significantly smaller than the variance of the control group? (The reasons for asking questions one and three are fully explained in TR 4.9: Design for Selection of Strategies and Media).

Table 43 presents the means and variances for the three major groups taking the final examination (experimental group, big control group, and pre- post control group). Table 44 presents the t-test comparisons and planned variance test comparisons on the final examination.

On the total test, while the mean of the experimental group was higher than that of the control group, it was not significantly so. On the total test, the experimental group answered constructed response items about as well as it answered multiple choice items. This condition did not obtain for the big control group, which answered multiple choice items significantly better ($p<.01$).

The variance test between the big control group and the experimental group indicated that the experimental group variance was significantly smaller ($p<.01$) than was the variance of the control group.

Audiovisual - Non-audiovisual Comparisons

For these comparisons three experimental groups, Audiovisual, Talking Book, and Illustrated Book were pooled to form the "AV groups". The other four experimental groups, Study Guide, Lecture Demonstration, Lecture, and Student Option were pooled to form the "non-AV groups". For each week t-tests

were conducted to test the equality of means for AV and non-AV pooled groups on media relevant items only. The seven weekly results are shown in Tables 45 and 46 for both total correct and mean log confidence scores. The total scores, that is total media relevant items for all AV groups were tested against total media relevant items for all non-AV groups. Since, in this case each subject was in all of the conditions, the resulting t-test has each subject as his own control. The results of this analysis are shown in Table 46, for mean log confidence scores only. The superiority of the non-AV groups seem clear.

Two of the criteria for selecting terminal objectives for which parallel media sequences[1] would be developed were the difficulty of the concept, and, the relative dependency of the concept on visualized motion for effective learning. A subset of media relevant items was specifically designed to compare motion-dependent versus motion-independent TO's. Another subset of test items was designed to compare difficult and non-difficult media relevant TO's.

In both cases, motion dependency and difficulty, it was expected that the parallel media groups would have superior performance on those specific test items. In neither case was this prediction substantiated by the data (both t's < 1.0).

Correlational Description of Variables

A correlational description of eighteen variables used

[1]See T.R. 4.9: Design for Selection of Strategies and Media for Elaboration, p. 25.

in the preceding analyses is given in Table 47.

The variables are SAT (M), QPR, Physics Validation, final examination score and total correct and log confidence average for each of the last seven weeks. Only those subjects which had complete data in all seven weeks (N = 77) were used.

Even though this description is not inferential it does point out the relatively high correlation between the total correct and the log confidence each week. The implication is that both scoring systems rank subjects in a very similar manner on performance.

The correlational data presented in Table 47 and 47a were collected for two principal reasons: to investigate the relationship between background variable scores and performance in the course, and, to obtain an estimate of the inter-relationships between performance measures, testing procedures, and rating procedures. Table 47a presents a different mix of background variables (than does Table 47) and, adds two additional performance variables, total media relevant test score, and total posttest score. Of particular interest is the relatively low correlation, .25 between experimental groups final examination performance and total posttest performance.

Study Guide Analysis

Tables 52 and 53 present the summarized data for the Study Guide Analysis. Since the Study Guide was the principal organizing feature of the course, a special analysis was made of student performance on the Study Guides. First,

it was desired to know whether student learning occured as a result of using the Study Guides.

Physics questions were presented to the student and he was asked to indicate his answer by marking a wet-to-reveal answer sheet. If his first choice was correct, he was directed to the next problem. If he was wrong, he was directed to a page with a remedial. He was then asked to answer the question correctly. If he was right, he was sent to the next problem; if wrong, he was sent to another remedial. He could miss each question a maximum of three times.

If pure chance were operating, the distribution should be as follows:

25% answer correctly on the first trial

33 1/3% of those remaining would answer correctly on the second trial

50% of those remaining would answer correctly on the third trial

100% would have answered correctly after the fourth trial

Chi-Squares with these theoretical frequencies were calculated for each of the Volumes A through O. The results of these Chi Squares are presented in Table 53. As can be seen, each of the Volumes had Chi Squares sufficiently high to reject the chance distribution hypothesis at well beyond the .01 level of confidence. Learning was clearly demonstrated.

When volume N was compared to the remaining volumes pooled, in an attempt to see if the different format of

volume N produced better results than the format for the remaining volumes, the Chi-Square value was 8.34, which fails to reach the established alpha level for a one-tail test. Analysis of Table 52 reveals that volume N might be slightly inferior, in that all volumes pooled had a .6 proportion passing for the three punch column, while volume N had only a .5 proportion passing.

Learning Category, Confidence, and Difficulty Ratings

When the project began, both Bloom and Gagnē conceptualizations of hierarchical orders of learning were analyzed to attempt to fit instruction appropriately to the categories. Finkel (1969) concluded that the instruction in the physics course was at the highest level in each of the systems mentioned. It was, therefore, decided that levels of problem difficulty, as rated by physicists, would be substituted for the hierarchies of learning.

The classification of questions was done according to the following scheme:

Learning Category (See examples in Appendix B)

0: Recognition of principles and recall of facts; substitution in formulas

1: Substitution into and solving of single step problems

2: Solving of multiple step problems

3: The "other" category into which questions were classified when they did not fit the other three categories.

Table 50 presents the intercorrelations, means, standard deviations, and N's for all test item rating data and

performance (mean proportion correct). Column 1 is mean proportion correct for the pooled experimental groups. Columns 2 and 3 are the faculty's item difficulty ratings in math and physics respectively. Column 4 is mean student recorded confidence, and column 5 is the students' mean difficulty rating. Column 6 is Learning Category, as described above.

Correlations between ratings and performance should, typically, be negative. The rating scales used make rated difficulty increase on five-point scales with 5 being most difficult. It was expected that a higher correlation between learning category and performance would have been found. While both students and faculty agree on the apparent difficulty of the test items, and, the relationships between learning category and the ratings are consistent and significant, (for 95 df, r for the .05 level is .17 and for the .01 level is .24), learning category and performance are not significantly related.

Table 51 presents the proportion passing test items of each learning category for the experimental groups and the pre-post control group. Generally, the proportion passing decreases as the nominal learning category increases, if one excludes category 3, the "other" category. The critical test is in the comparison of the pooled experimental groups and the pre-post control on media-relevant items. The performances are clearly not significantly different, and, even

ERIC

if they were, it is hard to imagine how such small differences would be important for decision making.

Both the pooled experimental and the pre-post control groups had higher proportions passing on media-relevant items than on non-relevant items, which might indicate that the media-relevant items were easier. Since the pre-post control did as well as the pooled experimental groups, and did not have the benefit of the media, an item difficulty differential appears reasonable, if it were not for the fact that the mean log confidence scores for both kinds of items shows a difference in the opposite direction. Correlational data (Table 47) indicate a high relationship between proportion correct and mean log confidence.

Analysis of Preference Data

Preference data were collected for two reasons: To find out which experimental conditions students preferred in general, and to obtain data on which revision decisions could be made.

A thirteen item checklist was developed for each media group. Each item was rated by the students on a five-point scale from "highly favorable" to "highly unfavorable". Ten of the thirteen items provided the student the chance to give a favorable or unfavorable reaction to the specific experimental condition and three items were designed specifically to check on unique features of each of the conditions. The three specific items were intended for revision purposes of the media, and were not classifiable as favorable or unfavorable.

The checklists were administered by the proctors in the various teaching areas on Monday of each of the last seven weeks of the semester. Each week, the students rated the experimental condition of the previous week.

Two separate analyses of the data were performed. First, ratings of all students were tallied for each of the experimental conditions across all seven weeks. These ratings were then combined across questionnaire items into one overall total for each of the experimental conditions. Table 48 presents the percentage of students, combined across weeks and items, in each of the experimental conditions responding to each choice on the five-point scale.

ERIC

The choices have been ordered from favorable to unfavorable, even though the scales were alternately reversed on the actual questionnaire to prevent position responding.

Two rank orderings were made of this data. First, columns 1 and 2 were combined into a single "favorable" proportion. Then, responses 4 and 5 were combined into a single "unfavorable" proportion. The result is a ranking of the experimental conditions in terms of "most favorable" to "least favorable" and a second ranking in terms of a "least unfavorable" to a "most unfavorable" order. Neutral responses were also ranked with the lowest proportion being assigned a rank of "1".

Tables 48 and 49 present the proportions of students responding by combined categories, the ranking of these proportions, and the rank ordering of time data.

The means and standard deviations were calculated for "favorable," "unfavorable" and "neutral" responses. The mean for favorable responses was .39 with a standard deviation of .084, unfavorable .19 and .053, and neutral .37 and .033.

The "Lecture" and "Student Option" conditions were essentially tied and were both more than one standard deviation above the mean, while the Lecture Demonstration condition was one standard deviation below the mean. These descriptions apply both to the favorable and to the unfavorable rankings.

The time data show a mean time in media of 171 minutes per week with a standard deviation of 61 minutes. Lecture and Student Option are essentially tied for the top rank and are both one standard deviation below mean time.

The rank correlation between time and preference data is .87, which allows rejection of the null hypothesis (rho = 0) at the .02 level.

The reported times indicated the amount of time actually spent by the student in the assigned rooms where each of the experimental conditions was applied. It was not required that students spend an equal amount of time on each of the experimental conditions. Specifically, no time was required of students under the "Student Option" condition. The times reported under student option indicate how much time these students spent in the experimental rooms when they were not required to do so.

V. DISCUSSION

The Effectiveness Report (Technical Report 5.6) describes the conclusions reached about overall course effectiveness. Generally, the conclusion was that the multi-media course was at least as effective, and probably more effective, than the traditional course. The basis for the claim that the course is as good as the traditional course is the virtual equality of control group and combined experimental mean scores on the final examination (Tables 43 and 44). The basis for the claim that the multi-media course is more effective lies in the significantly smaller variance of the experimental groups, as presented in Table 44.

Further, students tend to prefer the Student Option format of the course, the format recommended for final course implementation in the Management Systems Report (Technical Report 5.4) on an equal basis with the lecture. Table 49 presents the rank orderings of the preference data. Notice particularly that the student option and the lecture conditions are about equal in their preference ratings while the "LSG" condition (Lecture Demonstration) is least preferred. The definition of "traditional" instruction must include both Lecture and Lecture Demonstration. It is felt that there is a good logical basis for pooling the ratings for the lecture demonstration and lecture conditions to make a more thorough analysis of the preference data. No other experimental conditions can reasonably be pooled with the Student Option for preference purposes.

If one pools the Lecture and Lecture Demonstration Conditions, it seems clear that the Student Option condition is clearly preferred. While it is conceptually reasonable to pool Lecture and Lecture/Demonstration, these combinations were not specified _a priori_ and must be interpreted with considerable caution. (Data from the Spring and Fall of 1970 tend to indicate that there is a definate preference for Student Option over all other conditions. These data will be reported in detail in a supplement to this report as required by the contract modification of January 1971.)

Tables 23 and 24 present the data summaries and ANOVAS for the media relevant posttest items. These items, 30 in total, were selected before the beginning of the experiment to permit direct comparisons to be made among the various experimental treatments. These analyses were performed both on proportion correct and mean log confidence scores (see Technical Report 4.7, p. 28, for a discussion of the log confidence scoring system used). The F ratios for both log confidence and proportion correct were not sufficiently high to merit rejection of the null hypothesis. No evidence was found that the experimental conditions used were effective in producing differential performance. A further comparison was also designed, that of combining all "parallel media" groups into an audiovisual (AV) condition and the remaining experimental groups into a non-audiovisual (non-AV) condition. Each week, t-tests were made comparing the AV and non-AV conditions. Table 45 presents the

data on all of those t-tests conducted, with only one of 14 showing significance at even the .05 level of confidence.

These same media related items were further sub-divided into the media selection rationale categories described in the Design for Selection of Strategies and Media, Technical Report 4.9, p. 25. When the course objectives were written they were classified by the physicists as generally difficult, specifically difficult at the Naval Academy, and motion-dependent. There was a total of 158 terminal objectives, 27 of which were supplemented by the parallel media. One concern was to find out whether the students performed better on motion-dependent items when they had been instructed with videotapes. Further, the distinction could be made between motion-dependent difficult and motion-dependent non-difficult items. No significant differences were found in the comparisons between motion-dependent and motion-independent items, motion-dependent difficult and motion-dependent non-difficult items, motion-independent difficult and motion-independent non-difficult items.

While it is not surprising to find these results at the college level with a group of highly selected students, (Table 3 shows the SAT Verbal and Math scores for the experimental groups; the means for both math and verbal scores are above the 75th percentile of high school seniors who later enter college), it was felt that the techniques used for selection of the media would be effective in adding to the performance of students who used them. Regardless of the way that the data

was examined, it was not possible to discern any differential effects attributable to media. This statement is true, whether one maintains the integrity of a priori planned comparisons, or, does repeated ad hoc analyses by pooling groups on numerous bases in an attempt to tease out results.

When one examines student performance under all the experimental conditions employed in this tryout, it is difficult to find a basis for recommending the inclusion of specific audiovisual materials in the physics course. Apparently, the critical student responses, as viewed by physicists, are associated with problem working, as evidenced both by the final examination and the weekly posttests. Audiovisual (nonprint) media do not seem to improve the level of performance on these kinds of tests.

Analysis of the preference data adds little to the basis for recommending audiovisual media. The students said they preferred the Lecture and Student Option conditions and that they did not like the Lecture Demonstration condition. When questioned about this apparent discrepancy in preferences, a student explained simply: "The lecturer explains how to work problems similar to the ones which will be on the test. If I have to sit through a demonstration, then I don't learn how to work the problems.

"Under the Student Option condition, I can go to class if I want to, or, if I don't, I can learn how to work the problems from the other materials."

These results specifically led to the emphasis on the print media in the final version of the course. The Study Guides used in the Fall of 1969 were completely redesigned and revised on the basis of data collected during this tryout. These data and revisions are reported in Technical Report 5.5: Revision Process Documentation, particularly the discussion on p. 17ff.

Correlational Data

Table 47 presents the inter-correlations of the background variables and the performance variables. Each weekly posttest used the valid confidence scoring system described in Technical Report 4.7. The purpose of the valid confidence scoring system is to increase reliability of the test by asking a student to indicate his subjective probability of being correct on any test item. Columns 5 through 18 show the correlation between the weekly total correct scores and the confidence modified scores: week I, r = .98; week J, .88; week, .98, and so on.

These extremely high correlations between the confidence modified scores and the total correct scores would not seem to be a convincing argument for the addition of confidence modified scores to the simpler total correct scoring procedure. More clerical operations are required to convert the scores to the confidence scores, and, in addition, students are required to indicate their answer to the question as well as an estimate of subjective probability. Further, students typically resisted

the confidence approach until it was explained to them that they would receive no credit unless they complied.

Tables 21 and 22 present the results of the randomized block analysis. Significant media and weeks effects were found on both response measures, with log confidence achieving the established alpha level. The results are difficult to interpret, since only one of the weeks analyzed separately achieved even the .05 level. It is further difficult to see how media (experimental conditions) could have been a significant source of variation on total test items. (This analysis was performed using the total of media relevant and media non-relevant items.) Tables 35 and 36 present data for week L, and both log confidence and total correct reach the .05 level. Since this is the only week achieving the .05 level and, in light of the extremely high correlation between total correct and mean log confidence (Table 47), it is difficult to imagine how further attempts to interpret these results would be fruitful.

While the concept of learning category was not useful in this tryout in discriminating among students and kinds of test items, the contractor still views it as potentially useful for course development. The small number of items on which the appropriate classifications could be made may have had some bearing on the results. Further, because of the time constraint of the testing situation, it was not possible to control for the amount of time spent on each question or to specify in advance how much time should be allocated.

ERIC

Further investigation should occur in this area. Students appear to answer media relevant and media non-relevant items with equal facility (Table 53), independent of learning category (Mean log confidence for media relevant was 70.93, and for media non-relevant was 71.90). Perhaps experience with the items alone is enough to mask any differences that may exist, regardless of the type of instruction given. There appears to be no significant difference in the variances.

Why there should be the discrepancy in mean log confidence scores and proportion correct scores is unclear. Further investigation of this inconsistency is probably worthwhile in order to develop a reliable learning category classification system for physics problems.

It was not the purpose of the Fall 1969 tryout to arrange experimental conditions which would produce statistically significant differences. Rather, the purpose was to gather data which would be useful in revising the course to make it more appropriate for student-paced use. Procedurally, the kinds of data collected must be relatively inexpensive and require minimum time. Successive iterations of the course are not likely to improve student performance if such data cannot be used for purposes of revision.

It is one purpose of the course to increase mean student performance and reduce the variation in group performance. To that end, an examination of Table 44 indicates that progress was

made. While the difference between means favored the experimental groups, the difference was not statistically significant.

Not unexpectedly, all experimental groups did as well on those Final Examination questions requiring constructed response answers as they did on the multiple-choice questions. This was not true of the control groups, even though no correction for guessing was made.

The experimental conditions were all apparently equally effective in teaching students the required criterion behavior. It should be noted, however, that the criteria were based on a highly limited range of responses: the working of Physics problems. This conclusion seems warranted, regardless of whether one uses the norm-referenced Final Examination, or the criterion-referenced total posttest scores. In the special case of the media-related test items, the non-audiovisual groups did significantly better in total performance.

If one considers the performance data in light of the preference data, it appears that students are concerned with those experimental conditions which take the least time and which are most directly related to the content of the tests. For example, the L/SG condition was considerably less attractive to the students than was the straight Lecture (L) group. Conceivably, while the demonstration may have been interesting, the students viewed it as having no relationship to the important criteria of the course, namely, the working of Physics problems. While the inadequacy of such criteria has

been discussed more fully elsewhere (Branson, 1970) they are, nevertheless, widely used.

The preference of the SO condition may be attributable to the small amount of time actually prescribed for the students during those weeks. That is, if the lecturer is willing to show the student how to work problems, he is willing to listen. However, if one burdens the student with demonstrations, Audio-visual presentations, etc., the student seems much more willing to do it himself.

Regardless of the intergroup comparisons, the data collected are quite interesting. Each Terminal Objective was treated in a variety of ways: in the Study Guides, textbooks, and the lectures. The criterion-referenced test items used to measure the behavior were evaluated by the faculty along a number of dimensions: appropriateness to the TO (content validity), difficulty in Mathematics, difficulty in Physics.

These ratings are extremely valuable in providing a methodology by which a faculty member can, a priori, determine the level at which his course is taught. Provided that one is willing to accept final performance of the students as an indication of the level of sophistication of the course, the degree to which this can be specified in advance is a good indicator of the course "level."

If, on the other hand, it is necessary to wait until after the results are in to specify the level, it appears that the students, not the faculty, decide what performance is acceptable,

ERIC

particularly, if the grades in the course are assigned on any "normal" curve basis.

Our results indicate that the faculty and students are equally accurate in predicting student performance on the basis of difficulty ratings. Faculty correlations were -.43 and -.61 between performance and difficulty, while students' difficulty and performance correlated -.59.

This procedure for establishing course difficulty level appears imminently more desirable than a method which uses _ad hoc_ student performance to determine which test items should be retained and discarded. The results indicated a significant "weeks" effect, from which we inferred that weeks were not equally difficult. Physicists confronted with this data claimed to have known all along that some topics were indeed more difficult than others, as is virtually always the case in academic subjects.

The fact that the faculty could predict, with reasonable precision, the level of difficulty of the test items, and, thus, control this level of difficulty, transfers the responsibility of course level determination to the faculty.

The Study Guide results were of great general interest. While the "Linear-Branching" programmed instruction controversy has been dead for many years, it appeared reasonable in this course to offer specific remedial frames, to which the student was looped, when he failed to answer correctly on the first attempt. Further, that more specific remedials would be more effective than general remedials. Volume N had "general"

remedials. That is, the remedial was simply a presentation of the correct way to work the problem. The remainder of the course used specific remedials. That is, each problem was analyzed and the most likely, common, and probable errors were selected for elaboration. The students were shown why they were wrong, not how to do the problem correctly.

If a remedial is effective, it ought to reduce the probability of error on the subsequent attempts at the answer. Thus, if a student has missed the correct answer on the first trial and is given a remedial, he ought to have a better chance to be right on the second attempt than someone not receiving the specific remedial.

On the basis of this data, it was decided not to include specific remedials dealing with student errors in subsequent versions of the course. Course developers would concentrate on a more careful description of the correct way of working the problems.

Finally, the very low correlation between the performance of students on the total of 159 criterion-referenced items and the 60 item norm-referenced Final is encouraging. Professors' judgment of performance on criterion-referenced items (-.43, -.61) is a better indicator of final score on these items than is total student performance on norm-referenced items (.25) used as a predictor. Since the posttest items had been carefully screened for content validity prior to their inclusion on the test, and had been judged according to their expected level of

difficulty, it was possible to make a more accurate determination of the actual course level of difficulty than would otherwise have been possible.

Subsequent versions of the course can use the Test Item Bank in a pretest form and establish a baseline of student performance, having available past performance on the same items as a comparison. It is important to note here that professor judgment, tempered by past experience, is the critical element in developing the criterion measures. Student performance alone is not used. Consequently, test items are not discarded when a large proportion of students answers them correctly. They are discarded when they are rated and judged inappropriate by the faculty.

The results of the Fall 1969 tryout demonstrated to the Physics' faculty that the method of instruction was not the critical element in student performance, an accomplishment of some magnitude. Further, that students could, when provided with the necessary instruction and materials, achieve good results on their own. And finally, that if data is collected systematically and used to revise the course components, improvements can be made at each successive iteration.

VI. STATISTICAL TABLES

TABLE 1

MEAN SCORES BY PERFORMANCE GROUP

(Control Group Subjects on Final Exam)

	GROUP		
	1 (Low)	2 (Med)	3 (High)
SAMPLE SIZE	58	54	54
1 SAT (V)	559.67	582.76	599.35
2 SAT (M)	642.59	660.29	691.11
3 H.S. Rank	522.05	562.54	582.41
4 Whole Man	57185.74	58048.43	60427.98
5 QPR	220.50	257.722	312.22
6 NAV	44.55	45.07	42.91
7 ACH	49.98	51.96	50.00
8 M-F	51.95	52.20	52.07
9 OCL	54.98	55.59	56.69
10 SIN	44.97	44.70	47.11
11 SPL	39.69	40.30	38.41
12 PHYS. Val.	21.88	28.15	30.22

TABLE 2

DISCRIMINANT ANALYSIS CLASSIFICATION CHECK MATRIX FOR FINAL EXAM

(Control Group Subjects on 12 Variables)

	CLASSIFIED GROUP			
ACTUAL GROUP	1 (Low)	2 (Med)	3 (High)	SAMPLE SIZE
1 (Low)	44	11	3	58
2 (Med)	15	27	12	54
3 (High)	3	14	37	54

ERIC

TABLE 3

MEAN SCORES BY PERFORMANCE GROUP

(Experimental Group Subjects on Final Exam)

	GROUP		
	1 (Low)	2 (Med)	3 (High)
SAMPLE SIZE	42	47	40
1	586.83	564.28	607.88
2	665.57	666.28	693.40
3	556.57	573.43	600.63
4	57775.85	59709.04	59814.03
5	234.76	282.36	316.45
6	45.02	44.19	47.60
7	50.00	46.94	51.98
8	50.52	53.79	53.35
9	55.67	52.79	58.40
10	45.07	47.64	45.65
11	38.36	35.09	39.40
12	24.71	24.62	31.45

TABLE 4

DISCRIMINANT ANALYSIS CLASSIFICATION CHECK MATRIX FOR FINAL EXAM

(Experimental Group Subjects on 12 Variables)

	CLASSIFIED GROUP			
ACTUAL GROUP	1 (Low)	2 (Med)	3 (High)	SAMPLE SIZE
1 (Low)	28	9	5	42
2 (Med)	10	32	5	47
3 (High)	3	7	30	40

ERIC

TABLE 5

MEAN SCORES BY PERFORMANCE GROUP

(Control Group Subjects on Final Exam)

	GROUP		
	1 (Low)	2 (Med)	3 (High)
SAMPLE SIZE	58	58	54
1 SAT (M)	642.58	660.29	691.11
2 QPR	220.50	257.72	312.22
3 PHYS. Val.	21.88	28.15	30.22

TABLE 6

DISCRIMINANT ANALYSIS CLASSIFICATION CHECK MATRIX FOR FINAL EXAM

(Control Group Subjects on 3 Variables)

	CLASSIFIED GROUP			
ACTUAL GROUP	1 (Low)	2 (Med)	3 (High)	SAMPLE SIZE
1 (Low)	45	6	7	58
2 (Med)	17	22	15	54
3 (High)	4	10	40	54

TABLE 7

MEAN SCORES BY PERFORMANCE GROUP

(Experimental Group Subjects on Final Exam)

	GROUP		
	1 (Low)	2 (Med)	3 (High)
SAMPLE SIZE	42	47	40
1 SAT (M)	665.57	666.28	693.40
2 QPR	234.76	282.36	316.45
3 PHYS. Val.	24.71	24.62	31.45

TABLE 8

DISCRIMINANT ANALYSIS CLASSIFICATION CHECK MATRIX FOR FINAL EXAM

(Experimental Group Subjects on 3 Variables)

	CLASSIFIED GROUP			
ACTUAL GROUP	1 (Low)	2 (Med)	3 (High)	SAMPLE SIZE
1 (Low)	27	12	3	42
2 (Med)	12	24	11	47
3 (High)	5	10	25	40

ERIC

TABLE 9

MEAN SCORES BY PERFORMANCE GROUP

(week J, log confidence, 12 Variables)

	GROUP		
	1 (Low)	2 (Med)	3 (High)
SAMPLE SIZE	34	51	43
1	595.79	586.65	573.72
2	666.44	680.00	674.21
3	580.88	577.45	575.05
4	58641.68	59361.80	59346.07
5	254.50	278.96	293.23
6	43.68	47.80	44.16
7	48.35	50.80	48.65
8	50.71	52.49	54.02
9	54.29	57.06	54.35
10	51.06	43.55	45.88
11	35.88	38.49	37.44
12	25.82	27.27	26.91

TABLE 10

DISCRIMINANT ANALYSIS CLASSIFICATION CHECK MATRIX

(week J. log confidence, 12 Variables)

	CLASSIFIED GROUP			
ACTUAL GROUP	1 (low)	2 (Med)	3 (High)	SAMPLE SIZE
1 (Low)	23	6	5	34
2 (Med)	12	24	15	51
3 (High)	6	11	26	43

TABLE 11

MEAN SCORES BY PERFORMANCE GROUP

(week N, log confidence, 12 Variables)

	GROUP		
	1 (Low)	2 (Med)	3 (High)
SAMPLE SIZE	32	42	39
1	596.28	572.26	593.30
2	670.71	685.78	648.53
3	578.90	542.28	587.69
4	58892.12	58016.88	58403.02
5	252.18	272.02	297.07
6	44.93	45.40	45.74
7	47.78	50.47	63.43
8	52.18	52.16	62.46
9	54.46	55.85	54.12
10	48.75	44.00	51.35
11	36.12	37.04	37.15
12	25.65	27.73	24.69

TABLE 12

DISCRIMINANT ANALYSIS CLASSIFICATION CHECK MATRIX

(week N, log confidence, 12 Variables)

	CLASSIFIED GROUP			
ACTUAL GROUP	1 (Low)	2 (Med)	3 (High)	SAMPLE SIZE
1 (Low)	20	5	7	32
2 (Med)	8	26	8	42
3 (High)	5	10	24	39

TABLE 13

GENERALIZED MAHALANOBIS D^2 VALUES

SUBJECTS	GROUPING VARIABLE	NO. of VARIABLES	D.F.	D^2
Control	Final Exam	12	24	145.95**
Expt'l	Final Exam	12	24	135.31**
Control	Final Exam	3	6	124.43**
Expt'l	Final Exam	3	6	89.61**
Week I	Log Confidence	12	24	32.7
Week J	Log Confidence	12	24	51.8
Week K	Log Confidence	12	24	32.4
Week L	Log Confidence	12	24	35.7
Week M	Log Confidence	12	24	33.0
Week N	Log Confidence	12	24	62.2**
Week O	Log Confidence	12	24	72.6**
Week I	Proctor Time	12	24	23.9
Week J	Proctor Time	12	24	31.4
Week K	Proctor Time	12	24	31.9
Week L	Proctor Time	12	24	23.5
Week M	Proctor Time	12	24	33.7
Week N	Proctor Time	12	24	11.5
Week O	Proctor Time	12	24	19.6

TABLE 14

THE BEST REGRESSION PREDICTORS AND MULTIPLE

R by Week

Week	VARIABLE: TOTAL CORRECT	VARIABLE: LOG CONFIDENCE
I	SAT(V), QPR, M-F, SIN, (.58)	SAT(V), QPR, OCL, SIN, (.56)
J	QPR, (.31)	QPR, (.37)
K	NAV, Whole Man, (.32)	ACH, M-F, (.36)
L	Whole Man, M-F, Physics Validation, (.55)	Whole Man, M-F, Physics Validation, (.52)
M	SAT(M), QPR, (.47)	QPR, (.45)
N	QPR, SPL, Physics Validation, (.51)	QPR, OCL, SIN, (.52)
O	QPR, (.46)	QPR, (.43)

ERIC

TABLE 15

DATA SUMMARY AND ANOVA

Media Relevant Items - Total Correct

Weeks	I	J	K	L	M	N
Sample Size	140	150	150	144	137	130
Mean	3.3143	1.5267	.9000	.4583	1.0073	2.1462
Standard Deviation	.7206	.6625	.3010	.5000	.7225	.9970

Analysis of Variance

	Sum of Squares	DF	Mean Square	F Ratio
Between Weeks	757.3208	5	151.4642	331.5469**
Within Weeks	386.0305	845	.4568	
Total	1143.3514	850		

TABLE 16

DATA SUMMARY AND ANOVA

Media Relevant Items - Log Confidence

Weeks	I	J	K	L	M	N
Sample Size	140	150	150	144	137	130
Mean	83.743	74.750	90.247	54.208	57.730	59.968
Standard Deviation	17.501	33.402	29.202	42.160	31.683	21.229

Analysis of Variance

	Sum of Squares	DF	Mean Square	F Ratio
Between Weeks	160733.4398	5	32146.6880	34.6168**
Within Weeks	784704.2328	845	928.6441	
Total	945437.6726	850		

TABLE 17

DATA SUMMARY AND ANOVA

Media Non-relevant Items - Total Correct

Weeks	I	J	K	L	M	N	O
Sample Size	140	150	150	144	137	130	147
Mean	4.8714	4.2200	7.7800	6.2778	4.3212	2.3385	5.1156
Standard Devia - tion	1.1804	1.7869	1.1694	1.6702	1.4849	1.3210	2.1975

Analysis of Variance

	Sum of Squares	DF	Mean Square	F Ratio
Between Weeks	2471.9952	6	411.9992	163.1817**
Within Weeks	2502.0649	991	2.5248	
Total	4974.0601	997		

TABLE 18

DATA SUMMARY AND ANOVA

Media Non-relevant Items - Log Confidence

Weeks	I	J	K	L	M	N	O
Sample Size	140	150	149	144	137	130	147
Mean	61.949	44.410	71.272	59.381	48.306	37.012	45.859
Standard Devia - tion	13.721	16.596	10.117	14.137	13.198	14.542	19.420

Analysis of Variance

	Sum of Squares	DF	Mean Square	F Ratio
Between Weeks	121609.1115	6	20268.1852	92.4829**
Within Weeks	216964.5522	990	219.1561	
Total	338573.6637	996		

ERIC

TABLE 19

DATA SUMMARY AND ANOVA

Pooled Items - Total Correct

Weeks	I	J	K	L	M	N	O
Sample Size	140	150	150	144	137	130	147
Mean	8.1857	5.7467	8.6800	6.7361	5.3285	4.4846	5.1156
Standard Deviation	1.6164	2.1650	1.2549	1.7380	1.8235	1.9299	2.1975

Analysis of Variance

	Sum of Squares	DF	Mean Square	F Ratio
Between Weeks	2181.4164	6	363.5694	107.0440**
Within Weeks	3365.8792	991	3.3964	
Total	5547.2956	997		

TABLE 20

DATA SUMMARY AND ANOVA

Pooled Items - Log Confidence

Weeks	I	J	K	L	M	N	O
Sample Size	72	72	72	72	72	72	72
Mean	84.931	65.056	90.056	73.472	62.708	60.931	64.167
Standard Deviation	11.649	13.985	8.485	13.101	13.356	11.406	13.733

Analysis of Variance

	Sum of Squares	DF	Mean Square	F Ratio
Between Weeks	58521.000	6	9753.500	
Within Weeks	76111.000	497	153.141	63.690
Total	134632.000	503		

TABLE 21

ANOVA - WEEKS BY MEDIA BLOCK DESIGN

(Total Correct)

Source	SS	DF	MS	F Ratio
Media	43.53	6	7.26	2.09*
Weeks	2124.02	6	354.01	102.35**
Error	3257.99	935		
Total	5425.54	947	3.48	

TABLE 22

ANOVA - WEEKS BY MEDIA BLOCK DESIGN

(Log Confidence)

Source	SS	DF	MS	F Ratio
Media	3347.48	6	557.91	2.96**
Weeks	117593.18	6	19598.86	104.14
Error	187812.63	997	188.19	
Total	308753.29	1009		

TABLE 23

DATA SUMMARY AND ANOVA

Media Relevant Items - Total Correct

Media Group	A	B	C	D	E	F	G
Sample Size	120	124	109	129	133	116	119
Mean	1.5000	1.5323	1.5963	1.5194	1.4812	1.6293	1.5210
Standard Deviation	1.1739	1.1221	1.1395	1.1599	1.1781	1.1686	1.1778

Analysis of Variance

	Sum of Squares	DF	Mean Square	F Ratio
Between Media Groups	2.0234	6	.3372	.2504
Within Media Groups	1135.2719	843	1.3467	
Total	1137.2953	849		

TABLE 24

DATA SUMMARY AND ANOVA

Media Relevant Items - Log Confidence

Media Group	A	B	C	D	E	F	G
Sample Size	120	124	109	129	133	116	119
Mean	70.055	68.174	73.865	69.142	70.382	75.332	67.682
Standard Deviation	34.080	34.757	31.924	33.119	3[illegible].943	32.910	34.215

Analysis of Variance

	Sum of Squares	DF	Mean Square	F Ratio
Between Media Groups	5817.5554	6	969.5926	.8745
Within Media Groups	934645.8775	843	1108.7140	
Total	940463.4330	849		

TABLE 25

DATA SUMMARY AND ANOVA

Media Nonrelevant - Total Correct

Media Group	A	B	C	D	E	F	G
Sample Size	140	145	129	150	156	137	139
Mean	5.3643	4.9586	5.2946	4.5667	4.9487	5.1460	5.0863
Standard Deviation	2.2923	2.2107	2.3063	2.2298	2.2225	2.1711	2.1686

Analysis of Variance

	Sum of Squares	DF	Mean Square	F Ratio
Between Media Groups	60.7822	6	10.1304	2.0399
Within Media Groups	4911.4468	989	4.9661	
Total	4972.2289	995		

TABLE 26

DATA SUMMARY AND ANOVA

Media Nonrelevant - Log Confidence

Media Group	A	B	C	D	E	F	G
Sample Size	140	145	129	150	156	137	137
Mean	54.131	52.150	53.951	49.875	52.511	53.272	53.790
Standard Deviation	19.796	18.066	19.798	17.560	18.471	16.729	18.195

Analysis of Variance

	Sum of Squares	DF	Mean Square	F Ratio
Between Media Groups	2199.4086	6	366.5681	1.0847
Within Media Groups	333553.0820	987	337.9464	
Total	335752.4907	993		

ERIC

TABLE 27

DATA SUMMARY AND ANOVA

Pooled Items - Total Correct

Media Group	A	B	C	D	E	F	G
Sample Size	140	145	129	150	156	137	139
Mean	6.6500	6.2690	6.6434	5.8733	6.2115	6.5255	6.3885
Standard Deviation	2.4668	2.3489	2.3479	2.3553	2.4521	2.2657	2.1952

Analysis of Variance

	Sum of Squares	DF	Mean Square	F Ratio
Between Media Group	66.1386	6	11.0231	1.9938
Within Media Groups	5467.7520	989	5.5286	
Total	5533.8906	995		

TABLE 28

DATA SUMMARY AND ANOVA

Pooled Items - Log Confidence

Media Group	A	B	C	D	E	F	G
Sample Size	145	149	132	151	154	140	140
Mean	71.800	68.852	72.265	66.517	69.675	71.364	70.407
Standard Deviation	18.323	17.344	17.353	17.683	18.077	17.370	16.421

Analysis of Variance

	Sum of Squares	DF	Mean Square	F Ratio
Between Media Groups	3471.9590	6	578.6598	1.8824
Within Media Groups	308633.3605	1004	307.4037	
Total	312105.3195	1010		

ERIC

TABLE 29

DATA SUMMARY AND ANOVA - TOTAL CORRECT

(Week I)

Media Group	A	B	C	D	E	F	G
Sample Size	20	20	18	20	24	17	20
Mean	7.9500	8.0000	8.3333	7.9500	8.4583	8.4706	8.0500
Standard Deviation	1.9861	2.1764	1.4552	1.3945	1.4136	1.5049	1.3169

Analysis of Variance

	Sum of Squares	DF	Mean Square	F Ratio
Between Media Groups	6.8125	6	1.1354	.4245
Within Media Groups	353.0436	132	2.6746	
Total	359.8561	138		

TABLE 30

DATA SUMMARY AND ANOVA

(Week I, Log Confidence)

Media Group	A	B	C	D	E	F	G
Sample Size	19	20	19	21	23	18	20
Mean	81.842	82.950	85.316	82.048	86.217	87.944	84.000
Standard Deviation	15.417	16.916	11.634	11.617	12.041	11.175	9.268

Analysis of Variance

	Sum of Squares	DF	Mean Square	F Ratio
Between Media Groups	602.7800	6	100.4633	.6130
Within Media Groups	21795.3914	133	163.8751	
Total	22398.1714	139		

TABLE 31

DATA SUMMARY AND ANOVA

(Week J, Total Correct)

Media Group	A	B	C	D	E	F	G
Sample Size	22	22	18	22	24	20	21
Mean	6.4091	5.3182	5.6667	5.3182	6.1250	6.1000	5.1905
Standard Deviation	2.1527	2.5145	2.1144	1.7563	1.6501	2.6931	2.1822

Analysis of Variance

	Sum of Squares	DF	Mean Square	F Ratio
Between Media Groups	30.2652	6	5.0442	1.0746
Within Media Groups	666.5267	142	4.6939	
Total	696.7919	148		

TABLE 32

DATA SUMMARY AND ANOVA

(Week J, Log Confidence)

Media Group	A	B	C	D	E	F	G
Sample Size	25	26	21	23	23	20	22
Mean	66.720	59.769	64.286	61.217	67.043	68.800	61.364
Standard Deviation	14.772	14.836	10.011	13.003	13.313	19.264	14.578

Analysis of Variance

	Sum of Squares	DF	Mean Square	F Ratio
Between Media Groups	1656.4984	6	276.0831	1.3251
Within Media Groups	31877.1016	153	208.3471	
Total	33533.6000	159		

TABLE 33

DATA SUMMARY AND ANOVA

(Week K, Total Correct)

Media Groups	A	B	C	D	E	F	G
Sample Size	21	23	19	23	22	21	21
Mean	8.9524	8.8696	8.6316	8.3913	8.5455	8.6667	8.7143
Standard Deviation	1.0235	1.0576	1.8016	1.1962	1.4050	1.1547	1.1464

Analysis of Variance

	Sum of Square	DF	Mean Square	F Ratio
Between Media Groups	4.7727	6	.7954	.4948
Within Media Groups	229.8673	143	1.6075	
Total	234.6400	149		

TABLE 34

DATA SUMMARY AND ANOVA

(Week K, Log Confidence)

Media Group	A	B	C	D	E	F	G
Sample Size	20	23	19	23	22	21	21
Mean	89.950	89.217	86.737	86.043	86.500	87.143	88.429
Standard Deviation	8.894	9.553	14.375	10.594	12.520	11.315	10.390

Analysis of Variance

	Sum of Squares	DF	Mean Square	F Ratio
Between Media Groups	284.2886	6	47.3814	.3799
Within Media Groups	17710.7181	142	124.7234	
Total	17995.0067	148		

TABLE 35

DATA SUMMARY AND ANOVA

(Week L, Total Correct)

Media Group	A	B	C	D	E	F	G
Sample Size	22	16	20	23	22	20	21
Mean	7.4091	6.4375	7.4000	5.8261	7.0000	6.4500	6.6190
Standard Deviation	1.8168	.7274	1.4654	2.1246	1.6619	1.8202	1.5961

Analysis of Variance

	Sum of Squares	DF	Mean Square	F Ratio
Between Media Groups	42.7098	6	7.1183	2.5053*
Within Media Groups	389.2624	137	2.8413	
Total	431.9722	143		

ERIC

TABLE 36

DATA SUMMARY AND ANOVA

(Week L, Log Confidence)

Media Group	A	B	C	D	E	F	G
Sample Size	24	16	20	23	22	22	21
Mean	74.875	67.562	77.750	64.652	74.864	69.000	70.571
Standard Deviation	16.791	8.189	12.806	16.859	13.625	14.703	13.204

Analysis of Variance

	Sum of Squares	DF	Mean Square	F Ratio
Between Media Groups	2784.2161	6	464.0360	2.2730*
Within Media Groups	28785.2637	141	204.1508	
Total	31569.4797	147		

TABLE 37

DATA SUMMARY AND ANOVA

(Week M, Total Correct)

Media Group	A	B	C	D	E	F	G
Sample Size	19	22	18	18	22	19	19
Mean	5.5789	5.5000	6.3333	4.7778	4.6364	5.3158	5.2632
Standard Deviation	2.0088	1.5040	1.8471	1.6290	1.7606	1.6348	2.1040

Analysis of Variance

	Sum of Squares	DF	Mean Square	F Ratio
Between Media Groups	36.0959	6	6.0160	1.8794
Within Media Groups	416.1231	130	3.2009	
Total	452.2190	136		

TABLE 38

DATA SUMMARY AND ANOVA

(Week M, Log Confidence)

Media Group	A	B	C	D	E	F	G
Sample Size	19	22	18	18	22	19	19
Mean	62.316	62.864	68.556	57.222	57.045	60.053	62.053
Standard Deviation	18.679	12.422	16.343	13.113	12.890	14.524	14.393

Analysis of Variance

	Sum of Squares	DF	Mean Square	F Ratio
Between Media Groups	1758.3734	6	293.0622	1.3607
Within Media Groups	27999.1010	130	215.3777	
Total	29757.4757	136		

TABLE 39

DATA SUMMARY AND ANOVA

(Week N, Total Correct)

Media Groups	A	B	C	D	E	F	G
Sample Size	16	21	15	22	20	19	17
Mean	4.5625	4.6667	4.4667	4.0000	3.5500	4.9474	5.4118
Standard Deviation	2.3372	1.9579	1.6847	1.6903	1.8771	2.0405	1.5435

Analysis of Variance

	Sum of Squares	DF	Mean Square	F Ratio
Between Media Groups	42.1167	6	7.0195	1.9696
Within Media Groups	438.3523	123	3.5638	
Total	480.4692	129		

TABLE 40

DATA SUMMARY AND ANOVA

([illegible] N, Log Confidence)

Media Group	A	B	C	D	E	F	G
Sample Size	16	21	15	22	20	19	17
Mean	60.125	58.762	57.000	52.818	51.350	59.579	62.294
Standard Deviation	15.832	14.195	12.479	12.408	11.684	14.151	11.240

Analysis of Variance

	Sum of Squares	DF	Mean Square	F Ratio
Between Media Groups	1844.0645	6	307.3441	1.7610
Within Media Groups	21467.5432	123	174.5329	
Total	23311.6077	129		

TABLE 41

DATA SUMMARY AND ANOVA

(Week 0, Total Correct)

Media Group	A	B	C	D	E	F	G
Sample Size	20	21	20	21	23	21	20
Mean	5.0500	5.0476	5.2500	4.5238	4.9130	5.8095	5.2000
Standard Deviation	2.5849	2.0119	2.6132	2.2720	2.2343	1.7498	1.9628

Analysis of Variance

	Sum of Squares	DF	Mean Square	F Ratio
Between Media Groups	19.0919	6	3.1820	.6455
Within Media Groups	685.1547	139	4.9292	
Total	704.2466	145		

TABLE 42

DATA SUMMARY AND ANOVA

(Week 0, Log Confidence)

Media Group	A	B	C	D	E	F	G
Sample Size	21	21	20	21	22	21	20
Mean	64.095	61.714	63.800	59.762	62.409	67.190	62.500
Standard Deviation	16.571	13.439	19.362	14.839	15.327	10.939	14.471

Analysis of Variance

	Sum of Squares	DF	Mean Square	F Ratio
Between Media Groups	673.7842	6	112.2974	.4888
Within Media Groups	31936.6610	139	229.7602	
Total	32610.4452	145		

TABLE 43

FINAL EXAM SUMMARY DATA

Variable and Group	Mean	Variance	Sample Size
Total Exam (Experimental)	36.20	48.09	146
Total Exam (Big Control)	35.17	75.96	189
Total Exam (Pre-post Control)	35.17	60.84	76
Subtest 1 (Experimental)	9.51	5.68	146
Subtest 1 (Big Control)	9.48	8.12	189
Subtest 1(Pre-post)	9.44	7.13	76
Subtest 2 (Experimental)	11.55	3.92	146
Subtest 2 (Big Control)	9.92	5.20	189
Subtest 2 (Pre-post)	11.10	4.33	76
Subtest 3 (Experimental)	8.94	5.93	146
Subtest 3 (Big Control)	9.18	8.58	189
Subtest 3 (Pre-post)	8.48	5.86	76
Subtest 4 (Experimental)	5.88	6.76	146
Subtest 4 (Big Control)	6.57	8.06	189
Subtest 4 (Pre-post)	5.84	6.25	76

TABLE 44

"t" - TEST AND VARIANCE TEST RESULTS FOR FINAL EXAM

	t-value
Total Examination	
Experimental vs. Big Control	.13
Experimental vs. Pre-post	.14
Subtest 1	
Experimental vs. Big Control	.01
Experimental vs. Pre-post	.03
Subtest 2	
Experimental vs. Big Control	6.86**
Experimental vs. Pre-post	1.57
Subtest 3	
Experimental vs. Big Control	-.79
Experimental vs. Pre-post	1.38
Subtest 4	
Experimental vs. Big Control	-2.28*
Experimental vs. Pre-post	.11
Total Examination	
Multiple Choice vs. Constructed Response	
Experimental	1.52
Big Control	4.15**
Variance Test	
$\frac{\text{Big Control}}{\text{Experimental}}$	1.58**

TABLE 45

AV VS. NON-AV t - TESTS

Percent Correct

	AV Group			Non-AV Group			
Week	Sample Size	Mean	Std. Dev.	Sample Size	Mean	Std. Dev.	"t" Statistic
I	33	84.64	8.7	39	85.18	13.8	-0.20
J	32	69.28	15.9	40	61.67	11.3	2.37
K	27	90.48	8.7	45	89.80	8.5	0.33
L	35	71.26	14.6	37	75.57	11.2	-1.40
M	32	63.75	12.1	40	61.88	14.4	0.59
N	33	62.30	12.2	39	59.77	10.7	0.94
O	24	64.13	13.9	48	64.19	13.8	-0.02

ERIC

TABLE 46

AV VS. NON-AV t - TESTS

Log Confidence

	AV Group			Non-AV Group			
Week	Sample Size	Mean	Std. Dev.	Sample Size	Mean	Std. Dev.	"t" Statistic
I	33	0.82	0.1	39	0.83	0.2	-0.37
J	32	0.62	0.2	40	0.54	0.2	1.73
K	27	0.90	0.1	45	0.89	0.1	0.32
L	35	0.67	0.2	37	0.72	0.1	-1.31
M	32	0.56	0.1	40	0.54	0.2	0.65
N	33	0.50	0.2	39	0.50	0.1	0.00
O	24	0.52	0.2	48	0.54	0.2	-0.37

TOTAL

AV x NON-AV: $t = -28.36$, 76 df, $p < .005$.

TABLE 47

CORRELATIONAL DESCRIPTION OF VARIABLES

ROW	SAT M	QPR	FINAL EXAM	PHY. VAL.	I CORRECT	I CONFID.	J CORRECT	J CONFID.
1	1.0000							
2	.2847	1.0000						
3	.3915	.6964	1.0000					
4	.3619	.3827	.5156	1.0000				
5	.0264	.4683	.2583	.0759	1.0000			
6	.0342	.4671	.2502	.0594	.9819	1.0000		
7	.0432	.3142	.1993	.0829	.3830	.3408	1.0000	
8	.0809	.3672	.2537	.1089	.4334	.4011	.8881	1.0000
9	.1673	.2161	.0980	.0722	.1300	.1135	.3061	.2569
10	.1710	.2291	.1305	.1188	.1229	.1030	.2962	.2737
11	.1979	.4265	.4756	.3718	.2066	.1817	.1058	.1871
12	.2525	.3891	.4263	.3505	.1845	.1615	.0794	.1609
13	-.0406	.4654	.3564	.0447	.2523	.2579	.1756	.2260
14	.0429	.4478	.3051	.0506	.2761	.2746	.1730	.2463
15	-.0174	.4084	.3472	-.0241	.3320	.3202	.2795	.2652
16	.0418	.4323	.3320	.0424	.3361	.3236	.2628	.2815
17	.0418	.4577	.4256	.1358	.3301	.3097	.2677	.1873
18	.0581	.4334	.3880	.1446	.2544	.2395	.2080	.1203

TABLE 47 (cont'd)

CORRELATIONAL DESCRIPTION OF VARIABLES

ROW	K CORRECT	K CONFID.	L CORRECT	L CONFID.	M CORRECT	M CONFID.	N CORRECT	N CONFID.
9	1.0000							
10	.9805	1.0000						
11	.1090	.1345	1.0000					
12	.0770	.1010	.9689	1.0000				
13	.1437	.1493	.2442	.2130	1.0000			
14	.1547	.1581	.2280	.2166	.9541	1.0000		
15	.2482	.2828	.1855	.1597	.4642	.3957	1.0000	
16	.2503	.2899	.1664	.1548	.4627	.4502	.9469	1.0000
17	.0795	.0922	.1993	.1449	.3394	.2571	.2866	.2137
18	.0673	.0693	.2031	.1701	.3343	.2913	.2372	.1983

ROW	O CORRECT	O CONFID.
17	1.0000	
18	.9368	1.0000

TABLE 47a

Intercorrelations of background and performation variables on those subjects from whom a complete set of data was available.

N = 77

	1	2	3	4	5	6	7	8
1. SAT Verbal								
2. SAT Math	.31							
3. Highschool Rank	.14	.06						
4. Whole Man	.17	.32	.57					
5. Quality Point Ratio	.22	.28	.43	.53				
6. Final Exam	.34	.39	.25	.27	.70			
7. Physics Validation	.23	.36	.20	.23	.38	.52		
8. Media Related	.05	.12	.03	.22	.28	.23	.04	
9. Final Posttest	-.03	.08	.09	.35	.40	.25	.11	.74

NOTE: - For 70 df, the .05 level is .23, the .01 level is .30.

TABLE 48

Percentage of total responses, by experimental condition, in each category of the rating scales. Column 1 is most favorable, Column 5 is least favorable. Data are combined across the last seven weeks of the semester and across all items on the questionnaire.

	1	2	3	4	5
Audiovisual	7	27	41	21	4
Talking Book	13	28	35	19	4
Illustrated Book	10	23	40	21	6
Study Guide	12	29	35	21	3
Lecture Demonstration	4	22	40	26	7
Lecture	23	27	32	13	4
Student Option	11	36	34	16	3

TABLE 49

Rank Ordering of Preference and Time Data. Rankings on percent most favorable, percent least unfavorable, percent neutral. Time data ranked from least to most.

	Favorable Proportion	Neutral Proportion	Unfavorable Proportion
AV	.34	.41	.25
TB	.42	.35	.23
IB	.33	.40	.27
SG	.41	.35	.24
LSG	.26	.40	.34
L	.50	.32	.18
SO	.47	.34	.19
$\overline{X}$ =	.39	.37	.242
S =	.084	.033	.053

	Rating Most F	Rank Least U.	Order Neu.	Time Rank Order Least to Most
L	1*	1**	1	2*
SO	2*	2**	2	1*
TB	3	3	3.5	4
SG	4	4	3.5	3
AV	5	5	6	6
IB	6	6	4.5	5
L/SG	7**	7*	4.5	7

Time
$\overline{X}$ = 171
S = 61

* = +1S
** = -1S

* = +1S
** = -1S

TABLE 50

Intercorrelations, means, standard deviations, and sample sizes for test item characteristics and student performance. Sample size for each coefficient is shown in parentheses under the coefficient. For 95 df, the .05 level is .17 and the .01 level is .24.

	Variable Numer					
	1	2	3	4	5	6
1. Performance (mean proportion correct)		-.43 (155)	-.61 (155)	.67 (136)	-.59 (107)	-.13 (138)
2. Faculty Mathematics Rating			.64 (155)	-.40 (136)	.47 (107)	.44 (138)
3. Faculty Physics Rating				-.50 (136)	.52 (107)	.24 (138)
4. Student Recorded Confidence					-.74 (107)	-.27 (120)
5. Student Difficulty Rating						.37 (95)
6. Learning Category						

Variable	Mean	Standard Deviation	Number of Items
1	.6178	.2459	155
2	2.2112	.8056	155
3	3.0234	.5418	155
4	82.9378	13.0649	136
5	3.0051	.7918	107
6	1.0652	.6857	138
7	7.6645	.7668	155

TABLE 51

Proportion passing for each learning category for media relevant and media non-relevant posttest items.

	Learning Category Media Relevant				Learning Category Media Non-Relevant			
	0	1	2	3	0	1	2	3
All Experimental Groups	.78	.74	.54	.66	.67	.62	.57	.61
Pre-Post Control	.84	.71	.53	.65	.69	.66	.60	.61

TABLE 52

Comparison of Volume N with all other Volumes Pooled. Proportion of correct answers on each trial.

ALL VOLUMES POOLED

One Punch		Three Punch	Four Punch	TOTAL
37819	16256	7209	4090	65374
	16256	7209	4090	27555
		729	4090	11299

Proportion Correct

One Punch	.5
Two Punch	.5
Three Punch	.6

VOLUME N

One Punch	Two Punch	Three Punch	Four Punch	TOTAL
1225	668	264	197	2354
	668	264	197	1129
		264	197	461

Proportion Correct

One Punch	.5
Two Punch	.5
Three Punch	.5

TABLE 53

Chi square values for each volume of the study guide. Observed frequencies represent multiple punches on the study guide answer sheet.

Vol.	Observed	Expected	df	x^2
A	1540782	1061	2	1452.2**
B	1269329	1029	2	1233.5**
C	563642	772	2	730.1**
D	825715	845	2	977.2**
E	753781	1060	2	711.1**
F	443253	646	2	686.2**
G	345693	618	2	559.4**
I	572342	656	2	872.5**
J	195306	422	2	462.8**
K	394206	604	2	652.7**
L	300950	603	2	499.1**
M	258121	576	2	448.1**
N	129850	377	2	344.4**
O	135505	477	2	282.9**

VII. REFERENCES

Anderson, T.W. An Introduction to Multivariate Statistical Analysis, New York: John Wiley and Sons, 1958

Branson, R.K. The criterion problem in programmed instruction. Educational Technology, 1970, X(7), 35-37.

Cohen, J.C. Statistical Power Analyses for the Behavioral Sciences. New York: Academic Press, 1969.

Cooley, W.W., & Lohnes , P.R. Multivariate Procedures for the Behavioral Sciences. New York: John Wiley and Sons, 1962.

Deterline, W.A. & Branson, R.K. Evaluation and Validation Design. (Technical Report No.4.7) Old Westbury, New York, New York Institute of Technology, 1969.

Deterline, W.A. & Branson, R.K. An Empirical Course Development Model. (Technical Report No.5.7) Old Westbury, New York, New York Institute of Technology, 1971.

Draper, N.R., & Smith H. Applied Regression Analysis. New York: John Wiley and Sons, 1966.

Dubin, R. & Taveggia, T.C. The Teaching-Learning Paradox. Eugene, Oregon: Center for the Advanced Study of Educational Administration, 1968.

Finkel, R. Rationale for Sequencing Objectives. (Technical Report No.3.5) Old Westbury, New York, New York Institute of Technology, 1969.

Hicks, Charles R. Fundamental Concepts in Design of Experiments. New York: Holt, Rinehart & Winston, 1964.

Markle, D.G. Final Report: The Development of the Bell System First Aid and Personal Safety Course. Palo Alto, California; American Institutes for Research, 1967.

Popham, W.J. & Husek, T.R. Implications of criterion referenced measurement. Journal of Educational Measurement, 1969, 6, 1-9.

VIII. APPENDICES

APPENDIX A

Probability of Group Membership:

All calculations of the probability of group membership by discriminant analysis is accomplished by the following procedure.

For the i^{th} group ($i = 1,2,3,\ldots\ldots, g$) the probability that a person will belong to the i^{th} group is given by

$$Pi = \frac{e^{(f_i - \max f_i)}}{\sum_{i=1}^{g} e^{(f_i - \max f_i)}}$$

Where

g = number of groups used,

e = natural logrithm base,

$f_i = \sum_{k=1}^{v} z_k \quad C_{ki} - C_{oi}$,

C_{ki} = coefficients in the i^{th} column of the appropriate function table,

C_{oi} = constant for the same column above,

z_k = standard score on the k^{th} ($k = 1, 2, 3, \ldots\ldots, v$) variable for the person being classified, where v = number of variables, and

$\max f_i$ denotes the maximum value of all the f_i , $i = 1, 2, 3, \ldots\ldots, g$.

APPENDIX B

Description of Learning Category Taxonomy Used

The taxonomy consists of four categories:

1. Zero step questions -
   Those questions which require only the recall of a fact or definition, or the recognition of an object, fact, or definition. (see example 3-1 - Q2)

2. One step questions -
   (i) Those questions which require only direct substitution into an equation (usually algebraic) to be solved for one unknown. (see example 3-3 - Q6)

   (ii) Those questions which require correlation or association of two or more facts or definitions (but not directly requiring the facts or definitions for problem solution). (see example 3-3 - Q3)

   (iii) Those questions whose answers are a direct logical consequence of a fact of definition. (see example 3-3 - Q4)

3. Multiple step questions -
   All questions not falling into the zero- or one-step categories. (see example 3- Post Test - Q4)

4. Other -
   Those questions judged important by physicists, but not fitting into the other categories.

No distinction is made among two-, three-, or more-step problems for two reasons. First, the number of steps can be analyzed only into the intended behaviors, not the actual behaviors. Categorizing according to the above scheme minimizes the difference between intended and actual behavior. Secondly,

when more than a single operation (step) is required to solve a problem, even experts frequently disagree as to the "best" way to solve the problem and on what constitutes a "step." (Are intrinsic operations 'steps'?) Clearly, ambiguities in the step-counting process are much more likely to occur in multiple-step problems. Examples for Zero-, One-, and Multiple-step questions follow.

(Example 3-1-Q2)

2. "Uniform circular motion" refers to

A any circular motion.
B accelerating circular motion.
C circular motion without any acceleration.
D circular motion with constant speed.

(Example 3-3-Q6)

6. Near the surface of the moon, objects fall with an acceleration of 1.6 meter/sec^2. What is the weight of an object of mass 3 kg at the moon's surface?

A 4.8 nt.
B 2.8 nt.
C 1.8 nt.
D 3.8 nt.

(Example 3-3-Q3)

3. A rock weighs 64 lbs. on Earth. What does it weigh in free space, and what is its mass in free space? (the unit "slugs" is used as a shorthand notation for lbs. $sec.^2$ /ft, a unit of mass.

A weight in space 64 lbs, mass in space 0 slugs.
B weight in space 64 lbs, mass in space 2 slugs.
C weight in space 0 lbs, mass in space 64 lbs.
D weight in space 0 lbs, mass in space 2 slugs.

(Example 3-3-Q4)

4. The unit "newton" is a shorthand label for the units

A kg m/sec^2
B kg cm/sec
C kg sec/m
D kg sec^2/m

(Example 3-Post Test - Q4)

4 A light inextensible string is passed over a light, frictionless pulley. Two masses are suspended (vertically) from the ends of the string with mass m and the other with mass 2m. When the masses are released they have an acceleration

A g
B g/2
C g/3
D g/4
E 2g/3

ERIC